面向聚类的平面反射数据扰动方法*

汪小寒[a, b]，韩慧慧[a, b]，张泽培[a, b]，俞庆英[a, b]，郑孝遥[a, b]

(安徽师范大学 a. 计算机与信息学院; b. 网络与信息安全安徽省重点实验室, 安徽 芜湖 241003)

摘　要：面向聚类的数据隐藏通常使用数据扰动技术防止敏感信息泄露。针对现有的面向聚类的数据扰动方法隐私保护度低的问题，提出一种基于平面反射的数据扰动方法，将发布对象的全部属性两两配对构成平面上的点，再随机选择一条直线，作每对属性关于直线的对称点，转换后的数据即为发布的数据。实验结果表明，这种方法具有较好的隐私保护度和聚类可用性，且对高维数据有良好的适应性。

关键词：隐私保护；数据扰动；平面反射；聚类挖掘

中图分类号：TP309.2　　doi: 10.3969/j.issn.1001-3695.2018.01.0005

Planar reflection method of data perturbation for clustering

Wang Xiaohan[a, b], Han Huihui[a, b], Zhang Zepei[a, b], Yu Qingying[a, b], Zheng Xiaoyao[a, b]

(*a. School of Computer & Information, b. Anhui Provincial Key Laboratory of Network & Information Security Anhui Normal University, Wuhu Anhui* 241003, *China*)

Abstract: Data hiding for clustering usually uses data perturbation technology to prevent sensitive information disclosure. In order to solve the problem that the privacy protection of existing data-perturbation method for clustering is low, this paper proposes a data perturbation method based on plane reflection. All the properties of the published object are paired to form the points on the plane, then randomly select a straight line for each pair of symmetry points on the line, so the converted data is the data to be published. The experimental results show that this method has good privacy protection and clustering usability, and has good adaptability to high dimensional data.

Key words: privacy protection; data perturbation; plane reflection; clustering mining

0　引言

随着电子政务和电子商务的发展，个人数据在线交换不断增长，使数据收集变得越来越容易。数据挖掘者能从庞大的数据中提取许多有价值的信息，为广泛的应用提供支持。然而，这会引起个人隐私泄露和用户安全受到威胁的问题。因此，为了防止数据挖掘者收集大量隐私数据后泄露隐私信息，必须对原始数据进行处理。

数据匿名通过更改或模糊化原始数据方式更改或发布，改变后的数据即使与其他信息结合，也不能推理出任何关键信息。Sweeney[1]首先提出 *k*-匿名隐私保护模型，它的基本思想是发布用户信息的时候，用户的真实信息不能从 k-1 个用户中区分出来。后续基于 *k*-匿名模型的各种隐私改进原则，提出 *t*-closeness[2]，(*c*，*l*)-diversity[3]，Hybrid *k*-anonymity[4]等模型。并在不同的隐私保护领域广泛应用，如数据挖掘中[5]和位置服务中[6]。然而，经过匿名化处理后，会造成不同程度的信息损失，数据可用性降低。相比之下，扰动通过修改原始数据，并尽可能保留原始数据的特征，扰动后不仅保护了数据隐私，还能维持数据集的可用性，更适用于数据挖掘。

因此，数据扰动方法被广泛用于面向聚类的隐私保护中。现有面向聚类的扰动大多采用平移、缩放和旋转几何数据转换方法，存在几何数据变换函数是可逆的缺点，会导致处理后的数据隐私保护级别太低[7]。针对该问题，本文提出基于平面反射的数据扰动方法，不仅提高了隐私保护水平，还保留了良好的数据特征，更有利于聚类分析。

本文的主要贡献如下:a）提出一种新的数据扰动方法，利用平面反射改变原始数据，将发布对象的全部属性两两配对构成平面上的点，再随机选择一条直线，作每对属性关于直线的对称点，可以隐藏敏感信息，保护数据隐私;b）用理论证明了本文提出的方法是一种完全保距变换，扰动后的数据保持着良好的聚类可用性;c）采用真实数据集，从隐私保护度和运行时间两个方面进行了实验与分析，结果表明提出方法隐私保护度较好，

收稿日期：2018-01-03；**修回日期：**2018-02-10　**基金项目：**国家自然科学基金资助项目（61702010，61772034）；安徽省自然科学基金资助项目（1708085MF156）；安徽师范大学创新基金资助项目（2017XJJ93）

作者简介：汪小寒（1978-），女，安徽枞阳人，副教授，硕士，主要研究方向为智能计算、信息安全（hanxiaoahnu@sina.com）；韩慧慧（1992-），女，硕士研究生，主要研究方向为信息安全；张泽培（1996-），男，硕士研究生，主要研究方向为信息安全.

对高维数据有良好的适应性和高效性。

1 相关工作

为了保护用户的敏感信息，研究者提出了不同的数据扰动方法。黄茂峰等人[8]利用对数螺线的几何性质，对原始数据进行扰动，能够有效的保护数据隐私，但是这种方法所选择的投影子集分割范围较小，对低维数据点进行扰动时，隐私保护强度不高。Guang 等人[9]利用奇异值分解方法，对不同的样品数据进行不同程度的扰动，代替原始数据，这种方法有效的平衡了数据隐私与数据的可用性，但只能用于分类问题，而数据挖掘是不限于分类的。为了提供云服务中的范围查询和 KNN 查询服务数据安全和保护，Xu 等人[10]提出一种随机空间扰动的数据扰动方法，然而，这种方法专门用于扰乱云服务中上传的数据，将用户的数据安全保存在云数据库中。

几何数据转换方法是面向聚类的数据扰动常用方法。Oliveira 等人[11, 12]利用平移、缩放和旋转转换方法，实现对数值数据聚类隐私保护。这种方法可用于低维数据和高维数据，算法的复杂度低，可扩展性强，但是隐私保护度较低。Rajalaxmi 等人[13]提出基于几何数据变换混合数据转换方法，对每个敏感属性从平移、缩放和旋转方法中随机选择两种进行操作，以满足隐私保护要求，保持一般聚类分析功能，但是这种方法主要解决分类中的聚类隐私保护问题。王静等人[14]提出了一种基于二次反射的数据转换方法(DRDP)，采用沿着对称轴反射的方法，得到新的点坐标，转换后的数据与原始数据相差较大，但是原始数据很容易被还原。Giannella 等人[15]指出当已知的原始数据元组少于数据维数时，攻击者只要有一组已知的原始数据元组（或输入），做很少的工作就能导致原始数据的泄露。

目前，国内外针对基于平面反射数据扰动方法的研究也取得了较好的效果。Achlioptas[16]提出基于随机映射的数据转换方法，通过乘以随机矩阵来扰动，达到隐私保护的目的，但是该方法聚类结果高度不确定，且只适用于高维数据的聚类。Oliveira 和 Zaïane[11]利用旋转的数据转换方法，将数据集的所有属性两两分组，当属性数目为奇数时，将剩余的那个属性与已扰动的某个属性组合在一起，每个属性对对应一个 n*2 的矩阵，将其乘以一个变换矩阵，从而实现对隐私数据的保护，但是这种方法攻击者只要利用矩阵的某些理论性质，就可以公开原始数据值。王静和汪晓刚[14]提出一种基于二次反射的数据转换方法，取敏感属性最大值和最小值之和的平均值向下取整得到的数值作为对称轴，沿着对称轴进行反射得到新的坐标点，但是这种方法原始数据很容易被还原。刘杰等[17]提出了一种基于平面反射几何数据转换方法，将发布对象的属性两两配对构成平面上的点，当属性数目为奇数时，将剩下的一个未配对的属性与一个已配对的属性进行配对，然后作关于一条直线的平面反射，但是它只对低维数据进行研究，并未涉及高维数据。

在许多情况下，攻击者没有任何先验知识似乎是不合理的假设，因此，现有的面向聚类扰动方法大多数具有隐私保护度低的缺点。本文提出的基于平面反射数据扰动方法，通过一些属性随机生成，所有属性随机两两配对，再经过任意一条直线的反射，对原始数据进行扰动，即使攻击者获得任意一组已知的原始数据元组，也不会导致原始数据的泄露，因此，本文提出的方法可以实现更高的隐私保护度。

2 预备知识

2.1 基本概念

定义 1 反射[18]。设 l 是平面上的一条定直线，平面上任意一点 p 关于 l 对称点为 p'。从点 p 以 l 为轴映射到另一点 p'，这种映射是平面上以 l 为轴的反射。

反射具有两个基本特征：a）点 p 与 p' 连线的中点在直线 l 上;b）点 p 与 p' 连线的斜率与直线 l 的斜率的乘积为-1。

定义 2 变换[18]。把平面中的每一个点，变成和它同一个平面内相应的唯一点，并且平面中的每一个点都是由相应的某一个点变换的，把这种平面中点的位置变化称作平面中一个点的变换。

定义 3 保距变换[18]。如果一个变换（记为 f ）把任意的两个点 A、B 变成 A'、B'，使 $d\left(f\left(A\right),f\left(B\right)\right)=d\left(A',B'\right)$，则这个变换具有保距性，称为保距变换。即经过变换，使任意两个原象之间的距离与转换后相对应的两个象之间的距离相等。

定理 1 从点 $P\left(X,Y\right)$ 到点 $p'\left(X',Y'\right)$ 的映射 δ，δ 在平面直角坐标系下表达式为

$$\begin{cases}X'=a_{11}X+a_{12}Y+b_1\\Y'=a_{21}X+a_{22}Y+b_2\end{cases}\quad(1)$$

其系数 $\begin{bmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{bmatrix}$ 是正交矩阵，那么映射 δ 为保距变换。

证明 设点 $M\left(X_1,Y_1\right)$，$N\left(X_2,Y_2\right)$ 是平面上的任意两点，他们关于映射 δ 的对称点分别为 $M'\left(X_1',Y_1'\right)$，$N'\left(X_2',Y_2'\right)$。

由式（1）可知：

$$\begin{bmatrix}X_1'-X_2'\\Y_1'-Y_2'\end{bmatrix}=\begin{bmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{bmatrix}\begin{bmatrix}X_1-X_2\\Y_1-Y_2\end{bmatrix}\quad(2)$$

由欧氏距离得：

$$\left|M'N'\right|^2=\left(X_1'-X_2',Y_1'-Y_2'\right)\begin{bmatrix}X_1'-X_2'\\Y_1'-Y_2'\end{bmatrix}\quad(3)$$

由式（2）（3）得

$$\begin{aligned}\left|M'N'\right|^2&=\left(X_1'-X_2',Y_1'-Y_2'\right)\begin{bmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{bmatrix}\begin{bmatrix}X_1-X_2\\Y_1-Y_2\end{bmatrix}\\&=\left(X_1-X_2,Y_1-Y_2\right)\begin{bmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{bmatrix}^T\begin{bmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{bmatrix}\begin{bmatrix}X_1-X_2\\Y_1-Y_2\end{bmatrix}\\&=\left|MN\right|^2\end{aligned}\quad(4)$$

因此，定理 1 中映射 δ 为保距变换。

定理 2 反射变换是保距变换。

证明 设直线方程为 $Y=kX+b$，任意一点 $M(X,Y)$ 关于直线的对称点为 $M'(X',Y')$。根据反射的两个基本特征得如下方程：

$$\begin{cases}\dfrac{Y+Y'}{2}=k\dfrac{X+X'}{2}+b\\ \dfrac{Y-Y'}{X-X'}=-\dfrac{1}{k}\end{cases} \tag{5}$$

由式（5）得

$$\begin{cases}X'=\dfrac{1-k^2}{1+k^2}X+\dfrac{2k}{1+k^2}Y-\dfrac{2kb}{1+k^2}\\ Y'=\dfrac{2k}{1+k^2}X+\dfrac{k^2-1}{1+k^2}Y+\dfrac{2b}{1+k^2}\end{cases} \tag{6}$$

因为 $\begin{bmatrix}\frac{1-k^2}{1+k^2} & \frac{2k}{1+k^2}\\ \frac{2k}{1+k^2} & \frac{k^2-1}{1+k^2}\end{bmatrix}\begin{bmatrix}\frac{1-k^2}{1+k^2} & \frac{2k}{1+k^2}\\ \frac{2k}{1+k^2} & \frac{k^2-1}{1+k^2}\end{bmatrix}^T=E$，所以矩阵 $\begin{bmatrix}\frac{1-k^2}{1+k^2} & \frac{2k}{1+k^2}\\ \frac{2k}{1+k^2} & \frac{k^2-1}{1+k^2}\end{bmatrix}$ 是正交矩阵，根据定理 1，该变换是保距变换。

2.2 聚类的可用性

在聚类挖掘时，通过分析聚类数据记录之间的相似性和聚类外部的相异性来划分聚簇。其中，在聚类分析中，距离是用来衡量数据记录之间的相似或相异的常用工具。如果一个数据集中的任意两个数据记录修改前后距离关系保持不变，就可以实现良好的聚类可用性。面向聚类的数据隐藏为了维持较高聚类可用性，需要保证修改前的数据集和修改后的数据集具有尽可能相似的聚簇结构和特征，并且应该保持数据集内的每个数据记录修改前后的聚簇标志尽可能不改变。本文提出的基于平面反射的数据扰动方法完全是一种保距变换，因此聚类可用性好，不影响数据挖掘结果的准确性。

2.3 几何数据转换方法

几何数据转换是基于图形成像原理，通过平移、缩放、旋转和反射等方式对原始数据进行扰动来隐藏敏感数值属性，同时保留原始属性的特征。

a)平移是指在同一个平面内，把一个图形上的所有位置坐标点沿着某一方向移动相同的距离，即对发布对象的每个敏感属性使用加法噪声扰动，使用的噪声项是常数且可以是正的或负的。

b)缩放变换常用于改变图形的尺寸，即对发布对象的每个敏感属性使用乘法噪声扰动，使用的噪声项也是常数且可以是正的或负的。

c)二维旋转是将图形沿着 xy 平面内的圆弧路径重新定位，即将发布对象的全部属性随机配对，配对后构成的点为旋转点的位置。噪声项是旋转角度 θ，旋转角的正值为绕旋转点的逆时针方向旋转，负值则是绕旋转点的顺时针旋转。

d)混合数据转换是指对发布对象的多个敏感属性任意选取平移、缩放和旋转等不同方式的变换，噪声项是常数或旋转角度 θ。

e)二次反射转换是指沿着对称轴反射的方法，获得新的坐标点。所谓二次反射是指点的横坐标和纵坐标都进行了反射，即属性对同时进行反射。设发布对象中共有 n 对敏感属性，OP_i 表示反射操作，$a_i(1\le i\le n)$ 是发布对象的任意敏感属性，对称轴取该 a_i 的最大值与最小值之和的平均值向下取整得到的数值。对敏感属性 a_i 进行 OP_i 操作，就是将 a_i 沿着对称轴进行反射。

下面把几何数据转换各种方法分别定义了统一噪声矢量，具体说明如表 1 所示。

表 1 定义统一噪声矢量

数据转换方法	统一噪声矢量
平移	（<add，常数 1>，<add，常数 2>）
缩放	（<mult，常数 1>，< mult，常数 2>）
旋转	（< rotate，θ >）
混合数据转换	（<mult，常数 1>，<add，常数 2>）
二次反射转换	（< a_i，OP_i >）

3 基于平面反射数据扰动方法

本文提出的基于平面反射数据扰动方法，通过把原始数据集的全部属性随机配对，然后任意选择一条直线作每对属性的对称点，以此对原始数据进行修改，从而把敏感信息隐藏起来，实现隐私保护。同时攻击者无法根据扰动后的数据恢复或重构出真实和完整的原始数据，但是从扰动后的数据中可以得到与原始数据聚类相同的信息，从而保持数据聚类可用性不变。

3.1 基本思路

首先将发布对象的全部属性两两配对构成平面上的点，再任意选择一条直线，作每对属性关于该直线的对称点，转换后的数据即为要发布的数据，具体方法概述如下：

a)将发布对象的所有属性两两配对构成平面上的点，如果发布对象所拥有的属性数目为偶数直接两两配对，否则随机生成一组与发布对象属性数目相同的数据，然后再两两配对。

b)手动随机设置直线的斜率和截距，产生一条直线，然后根据式（5）解得的等式（6）将配对后的每对属性作关于直线的对称点。

c)发布的数据即为转换后的数据。

3.2 算法实现

本文提出的基于平面反射数据扰动算法详细描述如下。其中，D_{m*n} 表示原始数据集，D'_{m*n} 表示转换后的数据集，m 表示原始数据集数据实例的个数，n 表示原始数据集属性的个数。设直线的斜率为 k，直线的截距为 b。

算法：基于平面反射的数据扰动方法

输入：原始数据集 D_{m*n}

输出：转换后的属性集 D_{m*n}'

1）get D_{m*n}；//读取原始数据集

2）If (n % 2 == 0)　//判断属性数目 n 是否为偶数

3）　temp_n = n；//将属性数目值暂存于 temp_n 中

4）Else

5）　temp_n=n+1；//属性数目加 1

6）end if

7）producepq ( )；//随机产生属性对

8）for each $D_{i*times} \in D_{m*temp_n}$ //每个数据实例

9）　If（times<temp_n）//随机生成一个数据

10)　　data.push_back ($D_{i*times}$)；

11)　　++times；

12）　Else

13)　　times=0；

14）　　data.push_back (rand ( ))；

15）　end if

16）　for each attributes（X，Y）//每对属性进行转换

17）　　$X' = \frac{1-k^2}{1+k^2}X + \frac{2k}{1+k^2}Y - \frac{2kb}{1+k^2}$；

18）　　$Y' = \frac{2k}{1+k^2}X + \frac{k^2-1}{1+k^2}Y + \frac{2b}{1+k^2}$；

19）　end for

20）end for

21）Output D_{m*n}'；

现举例说明上述的算法的转换结果，表 2 为原始数据，表 3 为转换后的数据。表 2 是由 2014 年全省县级常住人口调查主要数据公报得到的安徽省 16 个市县的部分区域数据。其中，Index 表示序列号，Area 表示每个市县的占地面积，Population 表示各个市县年末常住人口的数量，Population density 表示各个市县的人口密度。算法过程如下。

首先，1~7 行，由于发布对象的属性（Area、Population、Population density）数目为奇数，因此，将属性数目加 1，对属性进行编号，随机组成属性对，属性组的选取为：（Area，Population），(Population density，random attribute)；然后，8~15 行，读取表 2 中发布对象的每个数据实例，并随机生成一个属性数据；最后，16~20 行，按照选择的属性配对方式，随机选取 k=8，b=10 来对每对属性进行平面反射数据转换。转换后的数据如表 3 所示，21 行输出即为最终要发布的数据。

3.3 算法复杂度分析

在基于平面反射数据扰动方法中，步骤 a）时间复杂度为 $O(n)$，步骤 b）时间复杂度为 $O(m*n)$，所以总的时间复杂度为 $O(m*n)$。算法空间开销最大的是对发布对象的所有属性进行随机配对，需要开辟额外 $O(n)$ 的内存空间来存储配对好的数据，因此算法的空间复杂度为 $O(n)$。

表 2　原始数据

Index	Area（km^2）	Population（人）	Population density（人/km^2））
1	6911	7696000	1113.6
2	3317	3617000	1090.4
3	5952	3258000	547.4
4	2526	2375000	940.2
5	4049	2229000	550.5
6	2802	2159000	770.5
7	1113	738000	663.1
8	15329	5376000	350.7

表 3　转换后数据

Index	Area（km^2）	Population（人）	Population density （人/km^2）
1	1887700	7460900	3013.46
2	887121	3506520	1672.5
3	796198	3159220	2394.98
4	582165	2302550	533.16
5	544750	2161410	1395.3
6	528728	2093260	3282.99
7	180580	715567	7355.58
8	1308460	5214360	2963.72

4　实验与分析

实验环境为处理器 Intel(R)Core™ i5-6300HQ，内存 4GB，操作系统 Windows 10，程序编译环境 Visual Studio community 2015。实验的数据选自 DIM-sets(high)（http://cs.joensuu.fi/sipu/datasets/）中的聚类数据集，DIM-sets (high)是合成的数据，拥有 6 个高维数据集，维数分别是 32、64、128、256、512、1024，除了维数 256 的数据集数据数量是 1020，其他维数的数据数量均为 1024。实验结果至少测试 6 遍以上取测试平均值求得。主要从隐私保护度和运行时间两个方面来分析算法的性能。

4.1　隐私保护度

数据扰动后，可以通过计算原始属性值与扰动后的属性值之间的差异来评估隐私保护安全程度[19]，可以描述为：$S = Var\left(X - X'\right)/Var\left(X\right) = \sigma^2_{\left(X-X'\right)}/\sigma^2_X$（$\sigma$ 是方差函数）。其中，$Var(X - X')$ 的值越小，扰动前后的属性值越接近，反之扰动前后的属性值差别越大。

针对不同维度的数据，隐私度的定义为：$S' = \sum_{i=1}^{i=n} S_i / n$，其中，n 表示数据集维度个数，当数据集是单维度时，$S' = S$。S' 越大，扰动前后数据集的属性值差别越大，隐私保护安全度越高。

表 4 给出了当直线斜率 k 和直线截距 b 取值不同时，第 1

维、第 2 维和第 5 维数据的隐私保护度。

表 4　隐私保护度

序号	k	b	Var（第 1 维）	Var（第 2 维）	Var（第 5 维）
1	1	0	1.0020	1.3850	1.8901
2	1	43000	1.0020	1.3850	1.8362
3	1	-43000	1.0020	1.3850	2.0293
4	2	43000	2.5600	1.5097	2.0204
5	5	43000	3.6994	1.7621	1.8499
6	-5	43000	3.6985	2.2379	2.1788
7	50	43000	3.9968	1.9753	1.5996
8	-50	43000	3.9968	2.0244	2.2096
9	500	43000	0.0000	1.9976	1.7995
10	-500	43000	0.0000	2.0024	2.5574
11	50000	43000	4.0000	2.0000	1.8000

分析表 4，从 1~3 行可知，隐私保护度几乎不受 b 取值的任何影响，因为 b 值的改变相当于对直线作平移，即对原始属性值作平移。因此，b 值的随机选取有利于提高隐私保护度。由 7~10 行可以看出，k 取正值或负值对隐私保护度的影响不大。由 1、4、5、7、9 和 11 行可以看出，随着 k 值的增加，第 1 维的隐私保护度先增加再急速减少最后又急速增加，第 2 维的隐私保护度逐渐缓慢增加，第 5 维的隐私保护度围绕 1.8000 上下波动，由此可见，随着数据维数的增加数据隐私保护度适应性越好，即对高维数据有良好的适应性。

表 5 是几何数据转换各种方法，即平移（TDP）[12]、缩放（SDP）[12]、旋转（RDP）[12]、混合数据转换（HDP）[12]和基于二次反射的数据转换方法（DRDP）[14]，噪声矢量取不同值，第 1 维、第 2 维和第 5 维数据的隐私保护度的变化结果。

表 5　隐私保护度的比较

数据转换法	噪声矢量	Var（第 1 维）	Var（第 2 维）	Var（第 5 维）
TDP	(<add,5>,<add,-5>)	0	0	0
TDP	(<add,500>,<add,-500>)	0	0	0
SDP	(<mult,1.01>,<mult,0.99>)	0.0001	0.0001	0.0001
SDP	(<mult,100>,<mult,0.01>)	9801	4900.99	5580.992
RDP	(<rotate,50>)	0.0012	0.07	0.05786
RDP	(<rotate,-50>)	3.3822	3.6744	3.5896
HDP	(<mult,0.5> ,<add,2>)	0.25	0.125	0.15
HDP	(<mult,100>,<add,2>)	9801	4900.5	5880.6
DRDP	(< a_i， OP_i >)	4	4	4

分析表 5 可知，对于 TDP 方法，经过它变换的属性值看起来与原始属性值很相近，但隐私保护度几乎为 0。对于 SDP、RDP 和 HDP 方法，它们的隐私保护度会随着噪声矢量的选取，而急剧变化，因此隐私保护度不稳定，而 DPDR 方法采用二次反射的数据转换方法，在隐私保护度上有很大改进，但是原始数据很容易被还原。以上分析可知，几何数据转换的这些方法的隐私保护度非常低，其中 DPDR 方法在列出的五种方法中隐私保护度最好。列出的五中方法中，只要攻击者获得任意一个敏感属性数据，用户的敏感属性就会被泄露[15]，而本文方法所有属性两两配对具有随机性，b 值和 k 值的选取也是随机的，由于随机不确定性无法还原，即使攻击者获得任意一个敏感属性对，也不会泄露用户的全部敏感属性。表 4 也显示该方法基本不受 b 值和 k 取正值或负值干扰，且随着数据维度的增加，数据隐私保护度越稳定，本文提出的方法具有更好的隐私保护度。

4.2　配对方式对隐私保护度和运行时间的影响

本节验证发布对象的属性之间不同的配对方式对隐私保护度和运行时间的影响。选用属性为 6 和数据记录数量是 1000 的数据集，对所有属性随机配对会产生不同的配对方式，本次实验只选取了部分配对方式，实验结果如表 6 所示。

表 6　不同的配对方式下隐私保护度和运行时间的结果

序号	配对方式	隐私保护度	运行时间
1	(3,1)(4,2)(5,0)	2.4604	0.288
2	(0,1)(4,2)(5,3)	2.4414	0.288
3	(2,0)(4,3)(5,1)	4.4001	0.284
4	(1,3)(2,0)(5,4)	1.9493	0.284
5	(0,3)(4,2)(5,1)	2.438	0.282
6	(0,2)(1,5)(3,4)	4.7646	0.269
7	(0,2)(1,5)(3,4)	4.7646	0.280
8	(0,2)(1,4)(3,5)	3.2526	0.282

从表 6 可以看出，有七种配对方式且每种配对方式的隐私保护度都不相同，表明不同的配对方式隐私保护度不同。从 1 和 2 行、3 和 4 行，以及 5 和 8 行可知，不同的配对方式运行时间可能相等。从 6 和 7 行可知，即使配对方式相同，运行时间也可能不同。因此，算法的运行时间不受配对方式的影响，而隐私保护的强度却受配对方式的影响，同时配对方式的多样性，可以更好的保护隐私信息。

4.3　不同维数下运行时间的对比

为了验证维数对运行时间的影响，实验中分别选择 2、32、128、700、850 和 1024 维数据集，每个数据集分别在数据记录数量为 500、1000、1500、2000、2500 和 5000 下进行实验，算法的运行时间在不同维度下随着数据记录数量增加的比对结果如图 1 所示。

分析图 1 可知，当数据记录数量一定时，运行时间随着数据集维数增加而增加，曲线趋近于线性变化状态，时间复杂度较低，对高维数据具有良好的适应性，算法具有较好的稳定性。从图中可以看出，在 1024 维、5000 条记录的数据集下，运行时间是 160 多秒，算法性能很高。是由于发布对象的属性数目是偶数时，直接对每个属性进行编号并两两配对，读取属性数据和平面反射转换同步进行；当发布对象的属性数目是奇数时，

先将属性数目加 1，再对每个属性进行编号，然后对编号进行随机配对，其中，读取发布对象的属性数据和随机生成一个属性数据，以及每对属性进行平面反射数据转换均同步进行。

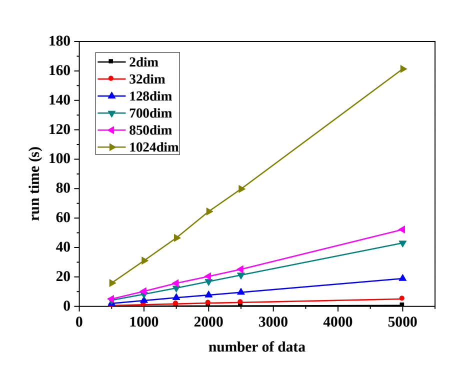


图 1　不同维数下运行时间比对结果

现有的多媒体数据、时间序列数据、空间数据、基因数据等聚类数据，常常会随着维数的增长，聚类的时间迅速增加导致算法性能下降，因此，本文算法更适合此类数据挖掘中对时间性能高要求的场景。

4.4　数据记录数量对运行时间的影响

本节验证了数据记录数量的变化对算法运行时间的影响。本次实验选用 6 维数据集，分别测试了数据记录数量是 500、1000、1500、2000、2500、3000、3500、4000、4500 和 5000 上的运行时间，并分别记录了每组数据多次测试的运行时间，然后结果取平均值，实验结果如图 2 所示。

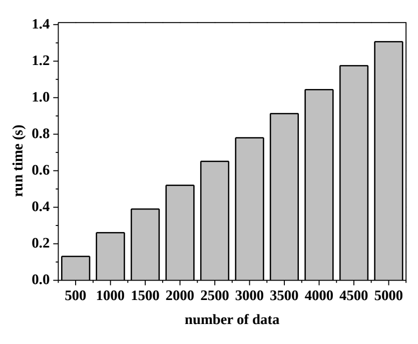


图 2　数据记录数量对运行时间的影响

分析图 2 可以看出，随着数据记录的增长，运行时间在不断增加，即数据记录越多运行时间越长，当数据记录为 5000 时，运行时间分别是数据记录为 500、1000 和 2500 的运行时间的 10 倍、5 倍和 2 倍，数据记录条目为 2000 时，运行时间分别是数据记录条目为 500 和 1000 的运行时间的 4 倍和 2 倍，由此可见，数据记录与运行时间成正比。同时表明，该算法能够处理大量的数据。同理，其他维数的数据集也呈现如此规律。

5　结束语

本文针对聚类的数据隐私保护问题，提出了一种基于平面反射数据扰动的方法，利用平面反射来干扰原始数据点的属性值，可以更好地保护数据隐私，保持聚类的数据可用性，使得数据挖掘结果准确度较高、算法的复杂性较低，且对高维数据有良好的适应性，但是该方法仅仅用于处理数值型属性，并没有对其他数据类型属性进行研究。

本文随机选择发布对象的所有属性两两配对，没有选取最优的配对方式，下一步将对提高隐私保护度进行更加深入的研究。

参考文献：

[1] Sweeney L. k-anonymity: a model for protecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10 (5): 557-570.

[2] Li N, Li T, Venkatasubramanian S. Closeness: a new privacy measure for data publishing [J]. IEEE Trans on Knowledge and Data Engineering, 2010, 22 (7): 943-956.

[3] Chakraborty S, Ambooken J G, Tripathy B K, *et al.* Analysis and performance enhancement to achieve recursive (c, l) diversity anonymization in social networks [J]. Trans on Data Privacy, 2015, 8 (2): 173-215.

[4] Aldeen Y A A S, Salleh M. A hybrid k-anonymity data relocation technique for privacy preserved data mining in cloud computing [J]. 인터넷정보학회논문지, 2016, 17 (5): 51-58.

[5] Bhaladhare P R, Jinwala D C. Novel approaches for privacy preserving data mining in k-anonymity model [J]. Journal of Information Science and Engineering, 2016, 32 (1): 63-78.

[6] Jia J, Zhang F. Nonexposure accurate location k-anonymity algorithm in LBS [J]. ScientificWorldJournal, 2014, 2014 (3): 619357.

[7] Banu R V, Nagaveni N. Evaluation of a perturbation-based technique for privacy preservation in a multi-party clustering scenario [J]. Information Sciences, 2013, 232 (5): 437-448.

[8] 黄茂峰，倪巍伟，王佳俊，等. 一种面向聚类的对数螺线数据扰动方法 [J]. 计算机学报, 2012, 35 (11): 2275-2282.

[9] Guang L I, Wang Y D. An improved privacy-preserving classification mining method based on singular value decomposition [J]. Acta Electronica Sinica, 2012, 9 (6): 529-534.

[10] Xu H, Guo S, Chen K. Building confidential and efficient query services in the cloud with RASP data perturbation [J]. IEEE Trans on Knowledge and Data Engineering, 2012, 26 (2): 322-335.

[11] Oliveira S R M, Zaïane O R. Achieving privacy preservation when sharing data for clustering [C]// Proc of Secure Data Management Workshop. 2004: 67-82.

[12] Oliveira S R M, Zaane O R, Agropecuaria E I. Privacy preserving clustering by data transformation [C]// Proc of Brazilian Symposium on Databases. 2003: 37-52.

[13] Rajalaxmi R R, Natarajan A M. An effective data transformation approach for privacy preserving clustering [J]. Journal of Computer Science, 2008, 4 (4): 320-326.

[14] 王静，汪晓刚. 一种新的保护原始数据隐私性的聚类算法 [C]// 第十

届中国科协年会论文集. 北京: 国防工业出版社, 2008: 7.
[15] Giannella C R, Liu K, Kargupta H. Breaching euclidean distance-preserving data perturbation using few known inputs [M]. [S. l.] : Elsevier Science Publishers, 2013.
[16] Achlioptas D. Database-friendly random projections [C]// Proc of the 20th ACM Sigmod-Sigact-Sigart Symposium on Principles of Database Systems. New York: ACM Press, 2001: 274-281
[17] 刘杰, 徐一凤, 张健沛, 等. 面向隐私保护聚类的平面反射数据扰动方法 [J]. 计算机工程与应用, 2013, 49 (6): 135-138.
[18] 丘维声. 解析几何 [M]. 北京: 北京大学出版社, 2015: 194-199.
[19] Lakshmi M N. Privacy preserving clustering by hybrid data transformation approach [J]. International Journal of Emerging Technology and Advanced Engineering, 2013, 3 (8): 2250-2459.